Sur la valeur d'une donnée

Thierry Berthier
Maître de Conférences en Mathématiques, Université de Limoges

Mai 2014 – Article n° IV.3

Un monde de données

Les données créées dans le monde seraient passées de 1.2 zettaoctets (un Zo = 10 puissance 21 octets) en 2010 à 1.8 Zo en 2011, 2.8 Zo en 2012 et devraient atteindre 40 Zo en 2020. On estime que le volume mondial de données double tous les 18 mois. A titre d'exemple, le réseau social Twitter produit quotidiennement 7 téraoctets de données (1To = 10 puissance 12 octets), Facebook engendre plus de 10 To chaque jour. Le grand radiotélescope Square Kilometre Array (SKA), qui sera opérationnel en 2024, produira plus d'un milliard de Gigaoctets de données par jour soit entre 300 et 1500 pétaoctets chaque année (1 Po = 10 puissance 15 octets). Le LHC, grand collisionneur de hadrons du CERN produit chaque année environ 15 Po de données. Le volume des données produites par les systèmes devrait rapidement dépasser celui produit par les humains.

Pour faire face à ce déluge de données, les technologies Big Data évoluent très rapidement selon trois axes désormais classiques, dits des « trois V » pour volume, variété, vélocité qui peuvent être complétés par deux autres V, la visibilité et la véracité. L'augmentation exponentielle des volumes de données à traiter induit la création de « Data Center » de plus en plus performants. La variété, qui traduit l'hétérogénéité des données brutes souvent peu structurées, est exploitable par une infrastructure algorithmique complexe capable d'interpréter l'information quel que soit son format. La vélocité quant à elle répond aux besoins de vitesses de traitements toujours plus élevées liées à l'analyse en temps réel des données (technologies « in memory ») et aux systèmes numériques « haute fréquence ». Le sixième V pourrait bien concerner la valeur d'une donnée, qu'elle soit liée à un Big Data ou non.

Existe-t-il une définition absolue de la valeur d'une donnée qui soit compatible avec l'environnement dans lequel elle est interprétée ou au contraire, doit-on se restreindre à une prise en compte relative, locale et instantanée ?

A partir de deux exemples récents, nous montrons que la valeur d'une donnée est avant tout une quantité volatile, temporelle et fortement dépendante du contexte sur lequel elle est évaluée. Nous proposons en seconde partie une approche systémique du problème, fondée sur un formalisme réduit qui permet de définir la valeur instantanée d'une donnée sur un contexte relativement à l'algorithme qui l'interprète.

I – Le prix d'une donnée selon deux exemples

1.1. Le tweet à 136 milliards de dollars

L' Armée Syrienne Électronique (SEA) [1], cellule cybercombattante apparue dès le début du conflit syrien, en 2011, apporte son soutien au régime de Bachar El Assad. Elle multiplie depuis trois ans les agressions numériques contre des cibles identifiées comme ennemi de la nation syrienne. Elle s'est donnée pour première mission de rétablir la vérité sur le conflit syriens notamment à partir d'une infrastructure structurée de contre-information déployée sur les réseaux sociaux (Facebook et Twitter), sur internet (site web sea.sy). La SEA a mené plus de 200 cyberattaques contre des intérêts numériques occidentaux de toute nature (médias, TV, grands journaux américains et européens, sites gouvernementaux européens, américains, arabes, israéliens, grands groupes comme Microsoft, Paypal, Facebook, Twitter, US Army,...).

Ses attaques reposent le plus souvent sur l'ingénierie sociale, l'intrusion (par phishing) et la prise de contrôle d'un compte à partir duquel l'opération est lancée. Les sites victimes sont régulièrement « défacés » par redirection vers une page similaire contenant un message de revendication et de justification de l'action. Quand le niveau de l'attaque le permet, la SEA procède à la captation de bases de données parfois très volumineuses. Ainsi, lors de l'agression menée contre le site Forbes en 2014, plus d'un million d'identifiants de comptes ont été piratés. L'attaque contre Paypal-UK avait permis de mettre la main sur une base de données du service de paiement en ligne.

La SEA utilise parfois l'attaque par déni de service (DDos) ou l'injection d'agents malveillants de collecte d'information plus sophistiqués en particulier contre la rébellion syrienne dans un objectif de renseignement.
Le 24 avril 2013, la SEA attaque le compte Twitter de l'agence Associated Press (AP). Elle en prend momentanément le contrôle et publie à 13h07 le message suivant :« *Breaking : Two Explosions in the White House and Barack Obama is injured* » soit : « *Deux explosions à la Maison Blanche, Barack Obama est blessé* ». Les 1.9 millions d'abonnés au compte Twitter d'Associated Press reçoivent le faux message posté par la SEA en le considérant comme authentique. La réaction des marchés financiers est presque immédiate : entre 13h08 et 13h10, l'indice principal de WallStreet, le Dow Jones (DJIA) perd 145 points soit l'équivalent de 136 milliards de dollars (105 milliards d'euros) en raison notamment du trading haute fréquence (HFT) qui a interprété et « réagi » au faux tweet. Les actions Microsoft, Apple, Mobil perdent plus de 1% presque instantanément. Quelques minutes plus tard, Associated Press reprend le contrôle de son compte et publie immédiatement un tweet annonçant que le message précédent était un faux et qu'il résultait du piratage de son compte. Aussitôt, l'indice Dow Jones remonte avec l'ensemble des valeurs qui venaient de chuter et reprend rapidement son cours normal. La courte durée de prise en compte du faux message publié par la SEA a suffi à modifier un indice boursier stratégique. L'activité des systèmes automatisés de trading haute fréquence, capables de passer des ordres en quelques microsecondes, a modifié les lignes de prise de décision repoussant le contrôle humain en fin d'opération.

La validation automatique et la prise en compte d'une information fausse peuvent donc avoir un impact considérable sur un environnement interconnecté. On peut alors s'interroger sur la valeur réelle du tweet SEA, en tant que donnée, considérée comme vraie à un instant puis démentie quelques minutes plus tard. Il est clair que cette valeur dépend à la fois de la variable temporelle mais également de la validation que l'on veut bien lui accorder et finalement du contexte sur lequel elle est interprétée. Il faut également s'entendre sur le sens du mot valeur : s'agit-il du meilleur prix de vente de cette donnée par un opérateur vers un opérateur acheteur ou doit-on tenir compte de la « valeur d'impact » de cette donnée sur un contexte ou sur un environnement plus global ? Dans le cas du faux tweet SEA, la valeur d'impact serait élevée puisqu'elle devrait prendre en compte le coût des turbulences exercées sur les marchés pendant la durée de validité de cette donnée.

1.2. La vente de données clients par Microsoft au FBI

En janvier 2014, l'hyperactive Armée Syrienne Électronique attaque à plusieurs reprises le site officiel de Microsoft et parvient à mettre la main sur plusieurs bases de données, courriers électroniques et factures

établies par Microsoft pour la vente de données « clients » au Bureau Fédéral d'Investigation (FBI). Le 21 janvier 2014, la SEA publie sur son site web la copie de nombreuses factures Microsoft envoyées au FBI ainsi que des listings de données personnelles vendues. Celles-ci concernent les utilisateurs d'Outlook ou de Skype et contiennent l'identité, l'identifiant, l'adresse IP, le nom de compte en hotmail.com et le mot de passe. D'après les factures publiées par la SEA, le coût unitaire d'un jeu de données concernant un utilisateur varie entre 50 dollars et 200 dollars en fonction du contenu transmis.

La facture pour le seul mois de novembre 2013 établie par Microsoft s'élève à 281 000 dollars. Celle du mois d'août 2013 atteint 352 000 dollars. Un jeu de données contenant le mot de passe de l'utilisateur du produit Microsoft est facturé 200 dollars (c'est le tarif maximal).

On notera que ce type de transaction est parfaitement légal dans la cadre d'une commission rogatoire intervenant pendant une enquête criminelle. D'autres factures sont établies par Microsoft auprès de sociétés étrangères privées basées en Amérique du sud dans le cadre de vente de données clients... Ces transactions viennent illustrer concrètement nos interrogations sur la valeur d'une donnée.

Dans le cadre d'une gestion massive de données, Microsoft parvient à définir un prix de donnée à l'unité, en fonction de son contenu et de son format. Dans ce cas, la valeur d'impact de la donnée n'est pas prise en compte par Microsoft dans l'élaboration du prix mais seulement son coût de traitement et de structuration.

Ces deux exemples induits par les cyberattaques de la SEA mettent en lumière la grande diversité des contextes valorisant les données et finalement, la réelle difficulté à proposer une définition canonique du prix d'une donnée. Une approche systémique peut contribuer à ordonner les paramètres et les composantes qui fondent la valeur de cette donnée.

II – Approche systémique

Nous proposons un formalisme minimal permettant de fixer une définition fonctionnelle de la valeur instantanée d'une donnée, sur un contexte, relativement à l'algorithme qui exploite cette donnée.

Définition 2.1. Donnée et mot binaire

- Une donnée est représentée par un ensemble fini de mots binaires.
- Un mot binaire est une suite binaire finie, c'est-à-dire une suite finie formée de 0 et 1, interprétable par un système de calcul.
- Cette définition permet de s'affranchir de la nature initiale d'une donnée (texte, image, son, vidéo, signaux ou mesures issus de capteurs,...). La totalité de l'information contenue dans la donnée initiale est traduite en mots binaires selon un format compatible avec un futur traitement algorithmique.

Notations 2.1

- On note D une donnée définie par $D = \{M_1, M_2, ..., M_n\}$ où les M_j sont des mots binaires, avec $M_j = b_1 b_2 .... b_k$ et $b_i = 0$ ou 1.
- vol(D) désigne le volume (en octets) de la donnée D non compressée. On parle aussi de taille de la donnée D. Lorsque que l'on compresse la donnée D à l'aide d'un algorithme de compression K, on note $vol_K(D)$ le volume de la donnée D après compression par K : $vol_K(D) = vol(K(D))$.

Définition 2.2. Contexte, sous-contexte et système

- On parlera de contexte pour désigner un ensemble d'infrastructures humaines, physiques et algorithmiques liées entre elles par des relations et des transferts d'information assurant une cohérence systémique globale. Un contexte est constitué de groupements humains et de

systèmes physiques et algorithmiques assurant son interconnexion.
- Tout sous-ensemble d'un contexte est appelé sous-contexte et peut être considéré comme un contexte plus restreint.

Exemples

Le marché international des matières premières est un contexte, celui du cacao un sous-contexte.
Le marché de l'art ou celui de l'énergie sont des contextes mondialisés. L'infrastructure de défense et de sécurité d'une nation est un contexte de niveau national. La Silicon Valley, en tant que pôle technologique, est un contexte local bien que fortement mondialisé. Un centre de données implanté en Silicon Valley est alors un sous-contexte de ce contexte.

Notations 2.2

- On notera C un contexte, S un système physique et algorithmique le composant et A un algorithme exécutable sur le système S du contexte C.
- Une donnée sera appréciée sur un contexte C, à un instant t, selon un algorithme A utilisé pour l'interpréter. La réunion de tous les contextes est notée Ω .
- Pour toute donnée D, à l'instant t, on a $\Omega = O_{D,t} \cup F_{D,t}$ (U étant le symbole de réunion) où $O_{D,t}$ désigne la réunion, à l'instant t, des contextes ayant accès au contenu informationnel de D et $F_{D,t}$,la réunion, à l'instant t, des contextes n'ayant pas accès au contenu de D.
- On définit une fonction indicatrice instantanée de la façon suivante :
  $I_{C,t}(D) = 1$ si le contexte C a accès à la donnée D, à l'instant t
  $I_{C,t}(D) = 0$ sinon.
  Ainsi, $O_{D,t} = \cup C$ tel que $I_{C,t}(D) = 1$ et $F_{D,t} = \cup C$ tel que $I_{C,t}(D) = 0$

- Une donnée est dite publique lorsqu'elle est connue et accessible à tous les contextes ($O_{D,t} = \Omega$ et $F_{D,t}$ est vide). Une donnée est dite privée, à l'instant t, si $F_{D,t}$ est non vide à l'instant t.
- Au cours du temps, une donnée privée peut devenir publique, mais pas l'inverse. L'ensemble $O_{D,t}$ est en général croissant avec le temps (au sens de l'inclusion) alors que $F_{D,t}$ n'est jamais croissant avec le temps car on suppose que l'information acquise ne se perd pas au fil du temps.

Définition 2.3. Valeur fonctionnelle instantanée d'une donnée sur un contexte selon un algorithme.

- Si D est une donnée accessible au contexte C, et A un algorithme interprétant D, exécutable sur un système de calcul S du contexte, on notera alors $Val_t(D / C, A)$ la valeur à l'instant t de D relativement au contexte C et à l'algorithme A exploitant D sur C.
- $Val_t(D / C, A)$ est une valeur numérique instantanée, positive ou nulle dépendant du contexte et de l'algorithme d'exploitation.

Définition 2.4. Valeur initiale d'une donnée sur un contexte.

- A l'instant initial t = 0, le contexte C prend connaissance pour la première fois du contenu informationnel de la donnée D et l'exploite selon l'algorithme A (qui peut n'être qu'un simple algorithme de lecture). Cette prise de connaissance du contenu de D résulte :
  - de la production de D par un composant de C qui la rend publique sur C
  - ou d'un simple achat de donnée vendue à C par une composante d'un autre contexte.
- $Val_0(D / C, A)$ désigne alors la valeur initiale de la donnée D sur C selon A. Elle peut être égale au prix d'achat par le contexte de la donnée D ou encore à son coût de production si le contexte à lui-même produit cette donnée.
- Lorsqu'une donnée D est publique, sa valeur instantanée peut ne pas être nulle, on parle alors de valeur résiduelle de la donnée : un contexte peut en effet avoir intérêt à acheter un jeu de données publiques préalablement structurées et raffinées par un autre contexte [2]. Le coût de raffinage et de structuration

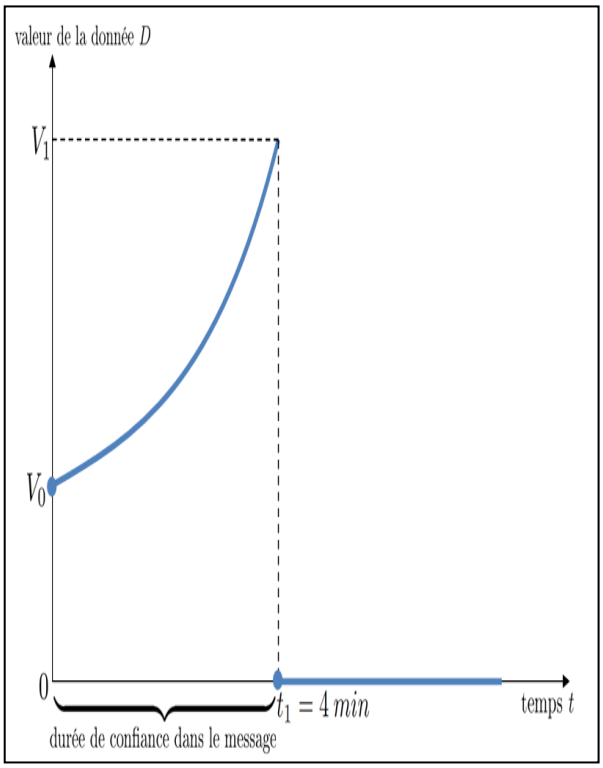
de la donnée publique engendre sa valeur sur d'autres contextes.

La nature de l'algorithme d'interprétation A de la donnée D sur le contexte C influence directement sa valeur instantanée [3]. Supposons par exemple que A désigne un algorithme qui commence par lire la donnée D puis calcule la probabilité p(D, t) que cette donnée soit vraie à l'instant t. Ce programme procède à un test de véracité sur la donnée avant qu'elle ne soit utilisée dans un environnement Big Data ou d'analyse sémantique. Si p(D,t) s'avère proche de zéro après calcul, la valeur de la donnée sera elle aussi proche de zéro dans un contexte rationnel. Si au contraire cette probabilité est proche de 1, le contexte considérera la donnée comme vraie ou presque vraie et pourra ensuite lui attribuer une valeur instantanée qui sera fonction de l'économie et des interactions entre contextes.

Lorsque D,C et A sont fixés, la fonction qui, à l'instant t, fait correspondre Val_t (D /C,A) décrit les variations instantanées de la valeur de la donnée D sur le contexte C selon l'algorithme d'interprétation A. Cette valeur évolue dans le temps, à partir d'une valeur initiale Val_0 (D /C,A) correspondant au coût de production de D sur C ou à son prix d'achat à l'instant t=0, jusqu'à sa valeur résiduelle notée Val_∞ (D /C,A). Une telle fonction peut présenter de fortes discontinuités comme le montre l'exemple 1 (fig1) du faux tweet créé par la SEA. Elle peut au contraire être constante sur le contexte qui l'a produite et structurée, comme dans l'exemple 2 (fig2) des données clients vendues par Microsoft au FBI pour 200 dollars l'unité.
La valeur instantanée dépend directement de la demande des contextes n'ayant pas encore accès à D et souhaitant l'acheter. Asymptotiquement, plus il existe de contextes qui connaissent le contenu de D et plus la valeur de D s'approche de sa valeur résiduelle. Quand la donnée devient publique (c'est-à-dire connue de tous les contextes), la valeur résiduelle est atteinte.

Fig.1 – Exemple 1 – La donnée est le faux tweet publié sur le compte d'AP

D = { Explosion à la Maison Blanche, le Président Obama est blessé }



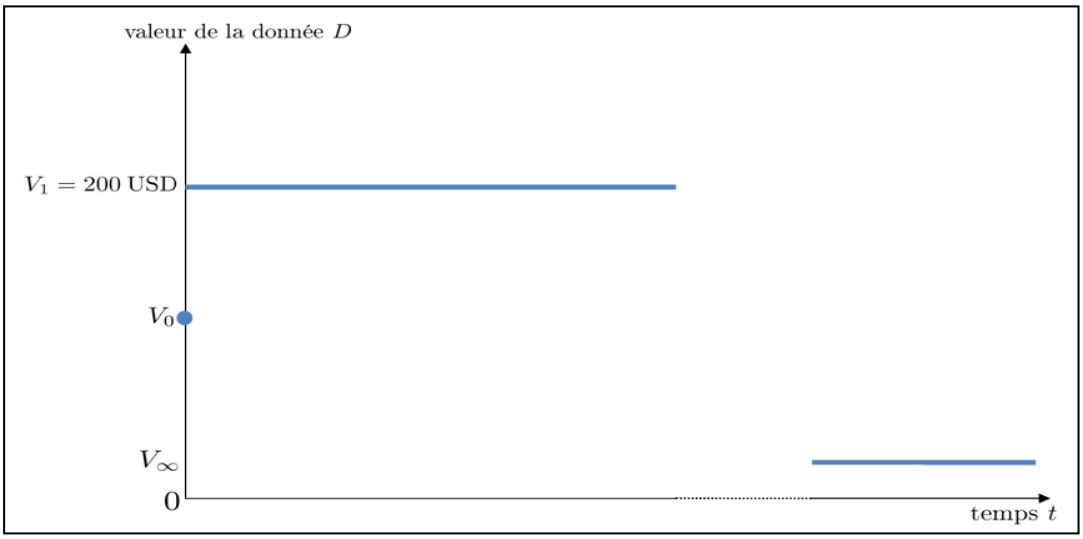
A l'instant t = 0, le tweet de la SEA est publié sur le compte AP et reste accessible et crédible durant quatre minutes. A l'instant t_1 , AP et la Maison Blanche publient un démenti qui annule immédiatement la valeur

instantanée de la donnée D.

V_0 désigne la valeur de production et d'insertion de la donnée sur le compte d'AP. Cette valeur tient compte du coût global du piratage du compte par la SEA.

V_1 est la valeur maximale de la donnée avant la reprise de contrôle du compte AP. Elle peut prendre en compte la valeur d'impact du faux tweet sur les marchés.

Fig.2 – Exemple 2 - Données client vendues par Microsoft au FBI



La valeur instantanée d'une donnée client D vendue par Microsoft au FBI vérifie :

- $Val_t(D / C, A) = 200$ USD pour $t > 0$ sur le contexte de production Microsoft.
- A est un algorithme de structuration (ou de mise au format) et de lecture de la donnée.
- V_0 est le coût de structuration, de mise au format et de stockage de la donnée pour Microsoft.
- V_1 désigne le prix de vente unitaire par Microsoft au FBI.
- V_∞ est la valeur résiduelle de la donnée.

Raffinage d'une donnée

Lorsque D, C, et t sont fixés, on dit que l'algorithme A' raffine la donnée D sur C à l'instant t mieux que l'algorithme A si : $Val_t(D / C, A') \geq Val_t(D / C, A)$
C'est le cas par exemple lorsque A se contente de lire la donnée D sur le contexte alors que A' lit cette donnée, évalue sa probabilité de véracité à l'instant t et montre qu'elle est proche de 1.
Le second algorithme apporte de la confiance sur la donnée, augmente donc sa valeur instantanée et raffine cette donnée sur C, mieux que A à l'instant t.
Au contraire, si la probabilité calculée par A' est proche de 0, alors $Val_t(D / C, A')$ sera proche de 0 et pourra dans ce cas être majorée par $Val_t(D / C, A)$ qui ne tient pas compte de la véracité de D.

Valeur instantanée et sous-contexte

Si C_1 est un sous-contexte de C_2, alors pour une donnée D fixée, pour un algorithme A et un instant t fixés, il

n'est en général pas possible de comparer $Val_t(D / C_1, A)$ et $Val_t(D / C_2, A)$. En effet, l'algorithme A peut se révéler plus efficace pour valoriser la donnée sur le sous-contexte ou au contraire sur le contexte plus étendu.

Origine et nature de la donnée

Lorsque la donnée D est engendrée sur le contexte C, on dit que C est son contexte d'origine. C'est le cas lorsque D est produite par un système de calcul installé sur C (objets connectés, systèmes de surveillance automatisés, instruments de mesures). Cette origine lui confère une nature systémique.
La donnée peut aussi être produite par un opérateur humain, à la suite d'une interaction avec un système de calcul. On parle alors de projection algorithmique d'un opérateur, selon un algorithme exécuté sur un système du contexte [4]. Dans ce cas, la donnée est dite projective.
Le volume global des données systémiques augmente aujourd'hui très rapidement et dépassera bientôt celui des données projectives. Quelle en sera la conséquence sur la valeur de ces données ? La donnée systémique sera-t-elle moins valorisée que la donnée projective ?

Diffusion de la donnée

Selon Philippe Davadie[1], le formalisme permettant de définir la valeur instantanée d'une donnée peut être complété par trois mesures effectives de diffusion de la donnée sur un contexte. Il propose d'introduire l'audience, l'écho et l'impact d'une donnée D.

L'audience instantanée d'une donnée D sur un contexte C, notée $Aud_t(D/C)$, mesure la fraction de la population des opérateurs du contexte C qui ont accès à la donnée D. C'est un nombre réel compris entre 0 et 1 qui vaut 1 si tout opérateur du contexte a accès à la donnée sans restriction particulière et 0 si, au contraire, aucun opérateur du contexte n'y a accès. Une donnée publique ouverte est par définition accessible à tout opérateur disposant d'un système de calcul interconnecté ; son audience vaut alors 1 sur tout contexte.

L'écho instantané d'une donnée D sur un contexte C, noté $Echo_t(D/C)$, mesure la fraction de la population des opérateurs du contexte C qui, ayant accès à cette donnée, l'utilisent réellement.

Enfin, l'impact instantané d'une donnée D sur un contexte C, noté $Imp_t(D/C)$, mesure l'effet de la donnée sur le contexte, c'est-à-dire sa capacité à modifier l'état du contexte, ses paramètres, ses réponses, et en définitive, son entropie. L'impact est certainement la quantité la plus difficile à cerner et à définir formellement en fonction des caractéristiques du contexte et de son interconnexion aux autres contextes. L'impact global de la donnée sur la réunion de tous les contextes détermine sa valeur d'impact. De même, l'audience et l'écho instantanés de la donnée influencent sa valeur instantanée. Existe-t-il alors des relations fonctionnelles simples entre ces trois mesures et la valeur instantanée d'une donnée ? Cette question reste ouverte.

Conclusion

Le formalisme introduit dans cette étude nous a permis de fixer les premières définitions de la valeur instantanée d'une donnée, sur un contexte, relativement à l'algorithme qui l'interprète. Cette approche relative et fonctionnelle s'écarte délibérément d'une description absolue tout en apportant une souplesse de représentation adaptée à la volatilité de la notion de valeur. Il reste à poursuivre son développement vers un modèle plus dynamique, équationnel ou non, capable de décrire les variations de la valeur d'une donnée pour mieux les anticiper. La donnée est une ressource, sachons en mesurer sa valeur !

[1] Philippe Davadie, Colonel, Centre d'Enseignement Supérieur de la Gendarmerie, est également membre des groupes de travail de la Chaire Cyberdéfense & Cybersécurité. Il vient de publier « *L'entreprise : nouveaux défis cyber* », éditions Economica, collection Cyberstratégie, 192 pages, mai 2014

Bibliographie

[1] Kempf Olivier et Berthier Thierry - « L'armée syrienne électronique : entre cyberagression et guerre de l'information » RDN – revue de la défense nationale – « Guerre de l'information » Vol. mai 2014.

[2] Janert Philipp K. - Data Analysis with Open Source Tools – O'Reilly.

[3] Bulusu Lakshman – Open Source – Data Warehousing and Business Intelligence – CRC Press

[4] Berthier Thierry - « Projections algorithmiques et cyberespace » R2IE – revue internationale d'intelligence économique – Vol 5-2 2013 pp. 179-195.

Sources en ligne

Site et compte Twitter Armée syrienne électronique :
https://twitter.com/Official_SEA16
http://sea.sy/index/en

Site de Simon Chignard sur l'Open Data :
http://donneesouvertes.info/

Site de la donnée publique :
http://www.data.gouv.fr/

Site de Thierry Berthier – cyberdéfense, cyberstratégies :
http://cyberland.centerblog.net/

Simulateur « Financial Times » calculant la valeur des données personnelles :
http://www.challenges.fr/entreprise/20130711.CHA2303/combien-valent-vos-donnees-personnelles-sur-internet.html
http://www.ft.com/cms/s/2/927ca86e-d29b-11e2-88ed-00144feab7de.html#axzz2WfFmKwic

Théorie de la valeur :
http://www.pise.info/eco/valeur.htm

Le prix des données publiques en France :
http://www.data-publica.com/content/2012/09/les-donnees-publiques-payantes-en-france-ce-quil-faut-retenir/#

Open Data France :
http://opendatafrance.net/category/donnees-publiques/

Données publiques – propriété intellectuelle – Sciences-Po
http://www.sciences-pi.fr/content/%C3%A0-qui-appartiennent-les-donn%C3%A9es-publiques

Site de Philippe Davadie : Informatiques orphelines et ouvrage à paraître- mai 2014 :
http://informatiques-orphelines.fr/

Fondation Saint-Cyr, Ecole militaire, 1 place Joffre, 75007 Paris
Téléphone: 01-45-55-43-56 - courriel: contact@chaire-cyber.fr; SIRET N° 497 802 645 000 18

La chaire remercie ses partenaires